Discovery of widespread GTP-binding motifs in genomic DNA and RNA

Published in final edited form as:
Chem Biol. 2013 April 18; 20(4): 521–532. doi:10.1016/j.chembiol.2013.02.015.

Discovery of Widespread GTP-Binding Motifs in Genomic DNA and RNA

Edward A. Curtis and **David R. Liu***
Department of Chemistry and Chemical Biology and Howard Hughes Medical Institute, Harvard University, 12 Oxford Street, Cambridge, MA 02138, United States of America

SUMMARY

Biological RNAs that bind small-molecules have been implicated in a variety of regulatory and catalytic processes. Inspired by these examples, we used *in vitro* selection to search a pool of genomeencoded RNA fragments for naturally occurring GTP aptamers. Several classes of aptamers were identified, including one ("the G motif") with a G-quadruplex structure. Further analysis revealed that most RNA and DNA G-quadruplexes bind GTP. The G motif is abundant in eukaryotes, and the human genome contains ~75,000 examples with dissociation constants comparable to the GTP concentration of a eukaryotic cell (~300 μM). G-quadruplexes play roles in diverse cellular processes, and our findings raise the possibility that GTP may play a role in the function of these elements. Consistent with this possibility, the sequence requirements of several classes of regulatory G-quadruplexes parallel those of GTP binding.

INTRODUCTION

Once thought to function primarily as a passive carrier of genetic information, RNA is now known to play an active role in diverse cellular processes (Tucker and Breaker, 2005; Huttenhofer et al., 2005; Bartel, 2009; Zhang et al., 2009). Although RNAs perform their biological functions using a variety of mechanisms, in a growing number of known cases binding cellular small molecules plays a critical role. Examples include riboswitches, which modulate gene expression in response to metabolites (Tucker and Breaker, 2005), and some ribozymes, which can require cofactors such as GTP and glucosamine-6-phosphate to catalyze reactions (Zhang et al., 2009). In most described cases, the role of RNA-bound small molecules is to modulate RNA folding in a manner that regulates gene expression (Tucker and Breaker, 2005). For example, in the presence of thiamine, the 5' untranslated region of the *E. coli thiC* mRNA adopts a secondary structure in which its Shine-Dalgarno sequence is inaccessible to the ribosome, resulting in decreased expression of downstream genes (Winkler et al., 2002). RNA-bound cofactors can also play catalytic roles in reactions catalyzed by ribozymes. The 3' hydroxyl group of an RNA-bound GTP molecule, for

*drliu@fas.harvard.edu, Fax: 617-496-5688, Phone: 617-496-1067.

AUTHOR CONTRIBUTIONS
E.A.C. and D.R.L. designed the research; E.A.C. performed the research and contributed new reagents/analytic tools; E.A.C. and D.R.L. analyzed the data and wrote the paper.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

example, acts as a nucleophile in the first step of the self-splicing reaction catalyzed by the Group I intron (Cech 1990).

Most examples of naturally occurring small molecule-binding RNAs have been identified using methods that search genomic databases for phylogenetically conserved RNA secondary structures (Barrick et al., 2004). This approach is especially suited to the discovery of riboswitches because they are often physically linked to the genes they regulate, greatly facilitating ligand identification (Winkler et al., 2002). Although a powerful way to identify riboswitches, bioinformatic methods are less applicable to aptamers whose ligands cannot be deduced from their genomic context, aptamers that lack canonical secondary structures, or aptamers whose degree of secondary structure conservation falls below a search threshold. A more general limitation of such methods is that they cannot be used to search for new motifs with biochemical functions specified by the experimenter.

Given these considerations, we speculated that additional examples of naturally occurring functional RNAs could be identified by directly selecting for motifs with specific biochemical activities. As an initial application of this approach, we used *in vitro* selection to search a pool of phylogenetically diverse genome-derived RNA fragments for new examples of naturally occurring GTP aptamers. GTP was chosen as our initial target because it is an essential molecule in all known organisms, it is widely used as a substrate by protein enzymes (Dever and Merrick, 1989; Alberts et al., 2007), and two known naturally occurring ribozymes use GTP as a cofactor (Cech, 1990; Teixeira et al., 2004). We identified several classes of genome-encoded GTP aptamers using this approach, the most abundant of which (the "G motif") adopts a G-quadruplex structure. Further analysis revealed that virtually all G-quadruplexes of both RNA and DNA bind GTP. We estimate that approximately 75,000 of the G-quadruplexes in the human genome bind GTP with a dissociation constant comparable to the GTP concentration of a typical eukaryotic cell, including members of several different classes of G-quadruplex regulatory elements. The sequence requirements of these elements parallel those of the G motif, suggesting that GTP binding may be involved in the function of some classes of regulatory G-quadruplexes.

RESULTS AND DISCUSSION

Construction of pools of genome-derived RNA fragments

To begin our search for new examples of naturally occurring small molecule-binding RNAs, we constructed pools of genome-derived RNA fragments from which such aptamers could be isolated (Figure 1a). Genomic DNA from several phylogenetically diverse eubacteria (*Escherichia coli*, *Bacillus subtilis*, and *Bacteroides fragilis*), archaebacteria (*Haloarcula marismortui*, *Aeropyrum pernix*, and *Methanococcus jannaschii*), and eukaryotes (*Homo sapiens* and *Gallus gallus*) was randomly fragmented using DNase I and fragments between ~100–600 bp were purified by gel electrophoresis.We chose genomic DNA as the starting point of these pools in light of studies suggesting that a larger fraction of the genome is transcribed than previously thought (Nielsen, 2011), and also because both low abundance and tissue-specific transcripts will be better represented in such a library than in one generated from cellular RNA.

These genomic DNA fragments were ligated into a vector, amplified by PCR, and transcribed using T7 RNA polymerase to generate pools of genome-derived RNA fragments flanked by defined primer binding sites suitable for *in vitro* selection experiments (see the Supplementary Information for detailed experimental procedures). In contrast with previous methods to select genomic RNA sequences with specific biochemical properties (Gold et al., 1997; Salehi-Ashtiani et al., 2006; Zimmermann et al., 2010), our pool contained genome-derived RNA fragments from multiple eukaryotic, eubacterial, and archaebacterial species.

Pools from different species can be mixed without sacrificing representation because the number of molecules present at the beginning of a typical *in vitro* selection experiment (~10^{15}) is far greater than the number needed to encode even a large genome.

Isolation of naturally occurring GTP aptamers by *in vitro* selection

The RNA pools (containing ~10^{12} molecules from each species) were combined and allowed to fold in a buffer containing 20 mM $MgCl_2$ and 200 mM KCl, similar to conditions in which naturally occurring aptamers and ribozymes are typically active. After incubating the combined pools with GTP immobilized through covalent attachment of its γ-phosphate to agarose beads (Sassanfar and Szostak, 1993; Connell and Yarus, 1994), bound molecules were eluted with EDTA, amplified by RT-PCR, and transcribed to generate RNA for the next round of selection (Figure 1b). After only two rounds of selection (and three GTP-agarose purifications; see the Supplementary Experimental Methods for details), GTP-binding activity of the enriched pool could be detected (Figure 1c), and after two additional rounds of selection, surviving molecules (the round 4 pool) were cloned and sequenced.

The GTP-binding activities of nine randomly selected clones from the round 4 pool were evaluated by comparing the fraction of each purified RNA that bound to GTP-agarose with the fraction that bound to control resin lacking GTP. Five of these nine clones bound at least 10-fold more efficiently to GTP agarose than to agarose beads lacking GTP, and four RNAs also bound at least 10-fold more efficiently to GTP agarose than did an unselected random N_{48} RNA pool (Table S1). Based on the number of GTP-agarose purifications needed to generate a pool with detectable GTP-binding activity, as well as the GTP-binding activities of individual aptamers isolated in the selection, we estimate that one in 10^3–10^5 genome-derived RNA fragments in the starting pool possess sufficient GTP-binding activity to survive the selection (see the Supplementary Experimental Methods for details).

Analysis of 73 unique sequences from the round 4 pool revealed that virtually all were derived from the genomes of either *Homo sapiens* (human, 42 sequences) or *Gallus gallus* (chicken, 25 sequences). Approximately half of these sequences mapped to intergenic regions, but others occurred in exons, introns, antisense to exons, or antisense to introns. To obtain more insight into potential roles of these GTP aptamers, we characterized their sequence requirements in greater detail.

Sequence requirements and biochemical characterization of the G motif aptamer

Initial examination of the sequences isolated in the selection revealed that virtually all were guanosine-rich (see the Supplementary Information for sequences), and that clusters of three or more consecutive guanosines occurred approximately 4-fold more frequently in both human and chicken sequences surviving the selection than they did in the genomes of these species. To better understand the relationship between G clusters and GTP-binding activity, we characterized the sequence requirements of one of these aptamers, clone 4–56, in greater detail. This aptamer was chosen for initial characterization efforts because (*i*) it is known to be expressed (as part of a 486-nucleotide transcript of unknown function) (Oh et al., 2005); (*ii*) pilot experiments revealed that its GTP-binding activity is conserved in primates (Figure S1), consistent with a possible biological role; and (*iii*) it is small (40 nts), which we anticipated would simplify its minimization.

The 4–56 aptamer contains four guanosine clusters separated by spacers, one of which has the potential to form a 13-nt hairpin (Figure S1). Deletion experiments revealed that this putative hairpin could be removed without loss of GTP-binding activity, that each of the three spacers linking the four guanosine clusters could be shortened to a single nucleotide, and that a seven-nucleotide minimized aptamer sequence containing only two guanosine

clusters separated by an adenosine spacer could also bind GTP (Figure 2a). Characterization of all 21 possible single-mutation variants of this minimized aptamer revealed that the identity of the spacer nucleotide separating the two G clusters is relatively unconstrained, but that point mutations at any of the other positions in the aptamer significantly reduced or completely abolished GTP-binding activity (Figure 2b). Further mutagenesis revealed that each guanosine cluster in the aptamer must contain at least three guanosines, that at least two guanosine clusters are required for efficient binding to GTP, and that in the context of adenosine spacers, as spacer length increases binding efficiency decreases (Figure 2c; also see Table S1). Taken together, these sequence requirements are consistent with those of a G-quadruplex structure (Davis, 2004) (Figure 2d). Consistent with this hypothesis, the circular dichroism (CD) spectrum of one of the most efficient GTP binders assayed is similar to that of previously described parallel strand G-quadruplex structures, with a positive peak at ~210 nm, a negative peak at ~240 nm, and a positive peak at ~260 nm [19] (Figure 2e). In contrast, CD spectra of RNA A-form helices contain a negative peak at ~210 nm, no peak at ~240 nm, and a positive peak at ~260 nm (Kypr et al., 2009).

Functional RNAs such as aptamers and ribozymes typically require metal ions for activity (Pyle, 2002). To characterize the metal ions requirements of the G motif, we first determined whether both Mg^{2+} and K^{+}, the metal ions present in the selection buffer, were required for GTP-binding activity using a construct that binds GTP efficiently by both GTP-agarose pulldown (Table S1) and gel filtration (Table S2) assays. Titration experiments revealed that in the absence of K^{+}, the G motif binds GTP optimally at ~3 mM Mg^{2+} (Figure S2a). In the absence of Mg^{2+} the aptamer can also bind GTP, with a sigmoidal dependence on K^{+} concentration and a plateau at about 1 M (Figure S2b). Binding efficiencies in Mg^{2+} alone and K^{+} alone were similar (Figure S2a–b), indicating that neither of these ions is absolutely required for aptamer function. To further characterize the metal ion requirements of the G motif, binding assays were performed separately in two additional monovalent ions (Li^{+} and Na^{+}) and six additional divalent ions (Ca^{2+}, Sr^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+} and Zn^{2+}). These experiments revealed that the G motif aptamer is active in a wide range of monovalent and divalent metal ions (Figure 3a). We note that although monovalent cations (especially K^{+}) can play important roles stabilizing G-quadruplex structures, numerous examples of G-quadruplex formation in the absence of monovalent metal ions have also been described (Chen, 1992; Smirnov and Shafer, 2000; Davis, 2004; Wei et al., 2008).

The affinity of the G motif for GTP was determined by measuring the amount of GTP bound by the aptamer as a function of GTP concentration (Figure 3b), revealing a dissociation constant (K_d) of 270 μM. The affinity of G motif for GTP is comparable to that of the Group I intron (Moran et al., 1993), a naturally occurring ribozyme that uses GTP as a cofactor, and is in a range that could be physiologically relevant in a eukaryotic cell (for example, the GTP concentration is 130 μM in HeLa cells (Finch et al., 1993), 300 μM in rat C6 glioma cells (Franklin and Twose, 1977), 340 μM in yeast (Koc et al., 2004), and 400 μM in *Xenopus* eggs (Woodland and Pestell, 1972)).

To characterize the binding specificity of the G motif aptamer for GTP, we determined the ability of sixteen GTP analogs to inhibit binding of radiolabeled GTP to the G motif aptamer using a gel filtration assay. GTPγS, GDP, GMP, guanosine, dGTP, ddGTP, and cGMP compete with GTP for binding to the G motif, while 7-deaza-dGTP, 7-methyl-GTP, 6-thio-GTP, 6-methylthio-GTP, ITP, XMP, ATP, CTP, and UTP do not (Figure 3c). These results suggest that the G motif makes contacts with both the Hoogsteen and Watson-Crick faces of GTP, but not with the hydroxyl or phosphate groups (Figures 3c and d), and are consistent with a model in which the G motif binds GTP by incorporating it as one of the guanosines in the tetrad of a G-quadruplex formed by the aptamer itself (Figure 2d).

Both RNA and DNA sequences are known to form G-quadruplex structures (Davis, 2004), and we hypothesized that DNA versions of the G motif might also bind GTP. To test this hypothesis, we assayed 30 DNA variants of the G motif (identical in sequence to those characterized as RNA sequences in Figures 2b and 2c) for their ability to bind GTP. The results revealed that RNA variants of the G motif that bind GTP are also typically active as DNA sequences, but that the GTP-binding activities of DNA variants of the G motif tend to be somewhat lower (3.7-fold lower on average in our assay) than their RNA counterparts (Figure 3e and Figure S2c–d).

The G motif aptamer is widespread in eukaryotic genomes

Both site-directed mutagenesis and CD experiments suggest that the G motif adopts a G-quadruplex structure. Based on this analysis, we hypothesized that the "quadruplex folding rule," previously shown to accurately predict sequences that can form intramolecular G-quadruplex structures (Todd et al., 2005; Huppert and Balasubramanian, 2005; Huppert and Balasubramanian, 2007), could also be used to determine the number of intramolecular examples of the G motif present in a dataset of interest (such as a sequenced genome). To test this hypothesis, we first used the quadparser algorithm (Huppert and Balasubramanian, 2005), which can identify various types of G-quadruplexes in sequence databases, to search each of the eight genomes from which our RNA pool was derived for the G-quadruplex consensus sequence $G_{(3–4)}N_{(1–7)}G_{(3–4)}N_{(1–7)}G_{(3–4)}N_{(1–7)}G_{(3–4)}$. This analysis revealed that the number of potential quadruplex forming sequences in these genomes ranged from six in *Bacillus subtilis* to ~140,000 in *Gallus gallus* (Table S3). Previous work has shown that ~376,000 G quadruplexes occur in the *Homo sapiens* genome (Todd et al., 2005; Huppert and Balasubramanian, 2005). When normalized for genome size, potential G-quadruplex forming sequences occurred most frequently in the *Aeropyrum pernix*, *Gallus gallus*, and *Homo sapiens* genomes (approximately once every 10^4 nucleotides), and least frequently in the genome of *Bacillus subtilis* (approximately once every 10^6 nucleotides) (Table S3).

To estimate the fraction of these potential G-quadruplex forming sequences that contain the G motif, 20 randomly chosen examples from the human genome were synthesized as DNA and tested for the ability to bind GTP. All 20 of these sequences bound GTP more efficiently than a random sequence control, and the GTP-binding activities of four out of the 20 matched or exceeded that of our reference G-motif (characterized in Figure 3) with a K_d of 270 μM (Figure 4a and Table S2). Consistent with this finding, the K_d of the construct with the highest GTP-binding activity (construct 20) was 100 μM (Figure 4b). To confirm that the tested sequences form G-quadruplexes, their structures were characterized by CD. Consistent with a previous test of the quadruplex folding rule (Huppert and Balasubramanian, 2007), all 20 sequences exhibited CD spectra consistent with G-quadruplex structures under the conditions tested (Figures 4c and S3).

Further analysis revealed two additional aspects of the relationship between G-quadruplexes and the G motif. First, the structures of the highest affinity aptamers identified in this study are parallel-strand G-quadruplexes, although not all parallel-strand structures bind GTP efficiently (Figures 4a and S3). Second, the sequences of the highest affinity aptamers typically contain short spacers connecting G clusters, consistent with the trends shown in Figures 2c and S2d. Indeed, sorting all tested sequences by maximum allowed spacer length using quadparser revealed that sequences encoding G-quadruplexes with spacers no longer than one nucleotide bound GTP more than 10-fold more efficiently than sequences encoding G-quadruplexes with longer spacers, and average GTP-binding activity continued to decrease as maximum allowed spacer length increased (Figure 4d and Table S4).

To more systematically characterize the phylogenetic distribution of the G motif aptamer, we used quadparser to search ~80 sequenced archaebacterial, eubacterial and eukaryotic

genomes for G-quadruplexes with spacers no longer than one nucleotide. This search revealed that the G motif is widespread in sequenced genomes, although its density (the number of examples of the G motif per nucleotide) varies by more than 1,000-fold (Figure 4e and Table S5). G motif density tends to increase with genome size, and is approximately 20-fold higher in eukaryotic genomes than in those of eubacteria or archaebacteria (Figure 4e and Table S5).

These results establish that the quadruplex folding rule can be used to identify new examples of the G motif in genomic sequence databases, and that, of the sequences that satisfy this rule, those with short spacers tend to bind GTP the most efficiently. In addition, they indicate that the density of the G motif is considerably higher in the genomes of eukaryotes than in those of archaea or eubacteria.

GTP-binding activity of G-quadruplex regulatory elements

G-quadruplexes have been shown to play roles in the regulation of diverse cellular processes (Kostadinov et al., 2006; Kendrick and Hurley, 2010; Bugaut and Balasubramanian, 2012). To investigate the possibility that GTP binding might play a role in the cellular function of G-quadruplex regulatory elements, we tested five well-studied examples for the ability to bind GTP: a transcriptional repressor in the human c-MYC promoter (Siddiqui-Jain et al., 2002; Seenisamy et al., 2004), an enhancer of 3’ end formation in the SV40 late transcript (Bagga et al., 1995), an internal ribosomal entry site (IRES) in the human VEGF gene (Morris et al., 2010), a translational repressor in the 5' UTR of the human NRAS gene (Kumari et al., 2008), and a G-quadruplex derived from the vertebrate telomere sequence (Blackburn, 2001). The GTP-binding activity of each of these regulatory elements was confirmed, although the telomere-derived sequence could only bind GTP in a buffer containing Sr^{2+}, which is known to promote the formation of parallel strand telomeric G-quadruplexes (Pedroso et al., 2007) (Figures 5a and S4). Mutations known to reduce cellular activities of these elements also significantly reduced their abilities to bind GTP (Figure 5a). Dissociation constants of these G-quadruplex regulatory elements for GTP ranged between 60 μM (for the c-MYC transcriptional repressor) and 600 μM (for the SV40 RNA processing enhancer; Figures 5b–e).

Although these results are consistent with the possibility that GTP binding plays a role in the cellular mechanism of regulatory G-quadruplexes, it is also possible that proper functioning of these elements simply requires that they adopt a G-quadruplex structure. Since not all G-quadruplexes bind GTP efficiently (Figures 4a and S3), one way to distinguish these possibilities is to compare the known sequence requirements of these regulatory elements with those of GTP binding. We performed this comparison using three different approaches. First, we investigated the extent to which the GTP-binding activity of these regulatory elements, rather than simply their ability to form G-quadruplex structures, has been conserved in evolution. This analysis indicated that, despite changes in primary sequence, the GTP-binding activity of the c-MYC transcriptional repressor and the VEGF IRES has been conserved in primates, and that of the NRAS translational terminator has been conserved in placental mammals (Figure 6).

Second, we investigated the correlation between the ability of previously characterized variants of G-quadruplex regulatory elements to perform their cellular function and to bind GTP. The ability of variants of both the c-MYC transcriptional repressor and the SV40 RNA processing enhancer to perform their cellular function was strongly correlated with their ability to bind GTP (Figure S5 and Table S1). In contrast, we observed only a weak correlation between the ability of variants of the VEGF IRES to promote cap-independent translation and to bind GTP (Figure S5).

Third, based on our observation that G-quadruplexes with short spacers bind GTP more efficiently than those with longer spacers (Figures 2c, 4d and S2d), we used bioinformatic methods to determine whether, for any of the genomic contexts in which G-quadruplex regulatory elements are known to be enriched, G-quadruplexes with short spacers show higher enrichment values than those with longer spacers. Enrichment values (defined as: density in a particular genomic context / density in the genome being examined) were determined for G-quadruplexes 100 bp upstream of human transcription start sites (Huppert and Balasubramanian, 2007), 100 nts downstream of human polyadenylation sites (Kostadinov et al., 2006), and in 5' untranslated regions of human mRNAs (Bugaut and Balasubramanian, 2012). G-quadruplexes with short spacers are enriched in each of these three contexts, and enrichment values show the expected decrease with increasing spacer length for G-quadruplexes upstream of transcription start sites (Table S6; see also Huppert and Balasubramanian, 2007). We also note that, although enrichment values of G-quadruplexes in human 5' UTRs do not appear to increase with decreasing spacer length, an example for which reducing spacer length increases translational inhibition has been reported (Halder et al., 2009).

Taken together, these results show that representative members of several different classes of regulatory G-quadruplexes bind GTP, with K_d values ranging between 60 and 600 μM. They also indicate that the sequence requirements of these elements, especially those involved in regulation of transcription, parallel the sequence requirements of the G motif.

Identification of a cellular protein that interacts with G motif-GTP complexes

More than 30 cellular proteins that interact with various types of G-quadruplex structures have been identified (Fry, 2007). These include nucleic acid binding proteins that promote the folding of G-quadruplexes, helicases that unwind G-quadruplex structures, and nucleases that specifically cleave phosphodiester bonds in G-quadruplexes (Fry, 2007). The existence of such factors led us to hypothesize that proteins that bind G motif-GTP complexes might also exist. Identification of such proteins would demonstrate that G motif-GTP complexes can interact with cellular components, and might provide clues about the potential biochemical roles played by these complexes in cells. To search for such proteins, we developed a gel filtration system that can separate biotinylated G motif-GTP-streptavidin ternary complexes from both biotinylated G motif-GTP binary complexes and free GTP molecules (Figure 7a). By performing this assay using unlabeled aptamer, unlabeled protein, and radiolabeled GTP, G motif-GTP-protein complexes can be readily distinguished from G motif-protein complexes of a similar molecular weight. We screened previously identified G-quadruplex binding proteins using this approach and observed that the zinc-finger protein CNBP (Calcaterra et al., 2010) forms a stable ternary complex with a DNA variant of the G motif and GTP (Figure 7b). Control experiments demonstrated that CNBP does not bind GTP by itself (Figure 7b), that G motif-GTP complexes do not nonspecifically interact with BSA (Figure 7b), and that the observed effect does not simply reflect CNBP increasing the amount of GTP bound by the G motif under these conditions (Figure 7c). No radiolabeled product was observed when these complexes were analyzed by either PAGE or SDS-PAGE, suggesting that neither GTP nor its gamma phosphate is covalently linked to either CNBP or the G motif.

CNBP is a 170 amino acid CCHC-type zinc finger protein that is highly conserved in vertebrates (Calcaterra et al., 2010). Originally discovered in a screen for proteins that bind the sterol regulatory element (Rajavashisth et al., 1989), CNBP also plays roles in forebrain development and cell proliferation, and has been linked to the human diseases myotonic dystrophy and sporadic inclusion body myositis (Calcaterra et al., 2010). Of particular interest to our studies, CNBP regulates transcription of the c-MYC gene by binding a G-rich region in its promoter that contains a phylogenetically conserved example of the G motif

(Figures 5a–b and 6a) (Michelotti et al., 1995). CNBP has also been shown to promote formation of parallel-strand G-quadruplex structures *in vitro*, suggesting that under certain conditions it might act as a chaperone for the G motif (Borgognone et al., 2010). Our finding that CNBP binds G motif-GTP complexes indicates that, of the many cellular proteins that interact with G-quadruplexes, at least one recognizes the same structural confirmation that binds GTP.

SIGNIFICANCE

RNAs play active roles in diverse cellular processes, and in an increasing number of known cases the ability to bind small molecules is an important aspect of their function. Inspired by these examples, we used *in vitro* selection to search a pool of genome-derived RNA fragments for aptamers that bind the biologically universal small molecule GTP. This revealed that DNA and RNA G-quadruplexes possess an intrinsic GTP-binding activity. The highest affinity binders form structures with parallel strands, suggesting that this is the optimal G-quadruplex topology for binding GTP. One way a G-quadruplex might bind GTP is by incorporating it into a tetrad of the G-quadruplex structure. Such a binding mode is consistent with the nucleotide specificity of the G motif elucidated in this work, as well as with previous observations that guanosine derivatives such as GMP can assembly into G-quadruplex structures when incubated at millimolar concentrations (Gellert et al., 1962).

Our study also indicates that GTP aptamers are abundant in eukaryotic genomes. For example, we estimate that ~75,000 of the G-quadruplexes in the human genome bind GTP with dissociation constants comparable to the GTP concentration of a typical eukaryotic cell, including motifs previously shown to regulate transcription, RNA processing and translation. This raises the possibility that GTP may play a role in the cellular function of these elements. Consistent with this hypothesis, the sequence requirements of several types of regulatory G-quadruplexes parallel those of GTP binding.

While this manuscript was in review a paper was published describing the discovery of an ATP aptamer in several bacterial and eukaryotic genomes (Vu et al., 2012). This motif was originally isolated from a random sequence pool (Sassanfar and Szostak, 1993) and had previously been identified in viral (Shu and Guo, 2003) and bacterial genomes (Laserson et al., 2005). A second recent study showed that the ydaO riboswitch binds ATP, and mediates gene expression in an ATP-dependent manner (Watson and Fedor, 2012).

EXPERIMENTAL PROCEDURES

See the Supplementary Information for detailed experimental procedures.

Pools were generated by fragmentation of genomic DNA using DNase I followed by gel purification of ~100–600 bp fragments on agarose gels. Fragments with 3' adenosine overhangs were generated by incubating first with DNA polymerase I, and then with dATP and *Taq* DNA polymerase. These fragments were ligated into pGEM-T vectors and amplified by PCR using primers flanking the insertion site, one of which contained a T7 promoter at its 5' end. Templates were transcribed using T7 RNA polymerase to generate starting pools for *in vitro* selection experiments.

GTP aptamers were isolated by incubating pool RNA with GTP agarose, washing away unbound molecules with selection buffer, and eluting bound RNAs with EDTA. Eluted molecules were subjected to RT-PCR and transcribed to generate RNA for the next round of selection. After four rounds of selection, the pool was cloned using the TOPO TA kit (Invitrogen) and sequenced.

The ability of aptamers to bind GTP was assessed in two ways: by measuring the amount of radiolabeled aptamer that bound GTP-agarose, and by determining the amount of radiolabeled GTP bound by the aptamer in a gel filtration assay.

Bioinformatic analysis of the G motif was performed using the quadparser algorithm. Analyzed datasets were obtained from either the NCBI website, the DOE Joint Genome Institute website or the UCSC Genome Browser. CD experiments were performed using a JASCO J-715 Spectropolarimeter.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Howard Hughes Medical Institute and the NIH/NIGMS (R01GM065865).

REFERENCES

Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, P. Molecular Biology of the Cell. New York: Garland; 2007.

Bagga PS, Ford LP, Chen F, Wilusz J. The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a trans-acting factor. Nucleic Acids Res. 1995; 23:1625–1631. [PubMed: 7784220]

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA. 2004; 101:6421–6426. [PubMed: 15096624]

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136:215–233. [PubMed: 19167326]

Blackburn EH. Switching and signaling at the telomere. Cell. 2001; 106:661–673. [PubMed: 11572773]

Borgognone M, Armas P, Calcaterra NB. Cellular nucleic-acid-binding protein, a transcriptional enhancer of c-Myc, promotes the formation of parallel G-quadruplexes. Biochem. J. 2010; 428:491–498. [PubMed: 20394585]

Bugaut A, Balasubramanian S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res. 2012; 40:4727–4741. [PubMed: 22351747]

Calcaterra NB, Armas P, Weiner AM, Borgognone M. CNBP: a multifunctional nucleic acid chaperone involved in cell death and proliferation control. IUBMB Life. 2010; 62:707–714. [PubMed: 20960530]

Cech TR. Nobel lecture. Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. Biosci. Rep. 1990; 10:239–261. [PubMed: 1699616]

Chen FM. Sr2+ facilitates intermolecular G-quadruplex formation of telomeric sequences. Biochemistry. 1992; 31:3769–3776. [PubMed: 1567831]

Connell GJ, Yarus M. RNAs with dual specificity and dual RNAs with similar specificity. Science. 1994; 264:1137–1141. [PubMed: 7513905]

Davis JT. G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. Angew. Chem. Int. Ed. Engl. 2004; 43:668–698. [PubMed: 14755695]

Dever, TE.; Merrick, WC. The Guanine-Nucleotide Binding Proteins. New York: Plenum Publishing Corp; 1989.

Finch RA, Revankar GR, Chan PK. Nucleolar localization of nucleophosmin/B23 requires GTP. J. Biol. Chem. 1993; 268:5823–5827. [PubMed: 8095498]

Franklin TJ, Twose PA. Reduction in beta-adrenergic response of cultured glioma cells following depletion of intracellular GTP. Eur. J. Biochem. 1977; 77:113–117. [PubMed: 198209]

Fry M. Tetraplex DNA and its interacting proteins. Front. Biosci. 2007; 12:4336–4351. [PubMed: 17485378]

Gellert M, Lipsett MN, Davies DR. Helix formation by guanylic acid. Proc. Natl. Acad. Sci. USA. 1962; 48:2013–2018. [PubMed: 13947099]

Gold L, Brown D, He Y, Shtatland T, Singer BS, Wu Y. From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. Proc. Natl. Acad. Sci. USA. 1997; 94:59–64. [PubMed: 8990161]

Halder K, Wieland M, Hartig JS. Predictable suppression of gene expression by 5'-UTR-based RNA quadruplexes. Nucleic Acids Res. 2009; 37:6811–6817. [PubMed: 19740765]

Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005; 33:2908–2916. [PubMed: 15914667]

Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. 2007; 35:406–413. [PubMed: 17169996]

Huttenhofer A, Schattner P, Polacek N. Non-coding RNAs: hope or hype? Trends Genet. 2005; 21:289–297. [PubMed: 15851066]

Kendrick S, Hurley LH. The role of G-quadruplex/i-motif secondary structures as cisacting regulatory elements. Pure Appl. Chem. 2010; 82:1609–1621. [PubMed: 21796223]

Koc A, Wheeler LJ, Mathews CK, Merrill GF. Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. J. Biol. Chem. 2004; 279:223–230. [PubMed: 14573610]

Kostadinov R, Malhotra N, Viotti M, Shine R, D'Antonio L, Bagga P. GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. Nucleic Acids Res. 2006; 34:D119–D124. [PubMed: 16381828]

Kumari S, Bugaut A, Balasubramanian S. Position and stability are determining factors for translation repression by an RNA G-quadruplex-forming sequence within the 5' UTR of the NRAS proto-oncogene. Biochemistry. 2008; 47:12664–12669. [PubMed: 18991403]

Kypr J, Kejnovska I, Renciuk D, Vorlickova M. Circular dichroism and conformational polymorphism of DNA. Nucleic Acids Res. 2009; 37:1713–1725. [PubMed: 19190094]

Laserson U, Gan HH, Schlick T. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. Nucleic Acids Res. 2005; 33:6057–6069. [PubMed: 16254081]

Michelotti EF, Tomonaga T, Krutzsch H, Levens D. Cellular nucleic acid binding protein regulates the CT element of the human c-myc protooncogene. J. Biol. Chem. 1995; 270:9494–9499. [PubMed: 7721877]

Moran S, Kierzek R, Turner DH. Binding of guanosine and 3' splice site analogues to a group I ribozyme: interactions with functional groups of guanosine and with additional nucleotides. Biochemistry. 1993; 32:5247–5256. [PubMed: 8494902]

Morris MJ, Negishi Y, Pazsint C, Schonhoft JD, Basu S. An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES. J. Am. Chem. Soc. 2010; 132:17831–17839. [PubMed: 21105704]

Nielsen H. The transcriptional landscape. Methods Mol. Biol. 2011; 703:3–14. [PubMed: 21125479]

Oh JH, Yang JO, Hahn Y, Kim MR, Byun SS, Jeon YJ, Kim JM, Song KS, Noh SM, Kim S, Yoo HS, Kim YS, Kim NS. Transcriptome analysis of human gastric cancer. Mamm. Genome. 2005; 16:942–954. [PubMed: 16341674]

Pedroso IM, Duarte LF, Yanez G, Baker AM, Fletcher TM. Induction of parallel human telomeric G-quadruplex structures by Sr(2+). Biochem. Biophys. Res. Commun. 2007; 358:298–303. [PubMed: 17485073]

Pyle AM. Metal ions in the structure and function of RNA. J. Biol. Inorg. Chem. 2002; 7:679–690. [PubMed: 12203005]

Rajavashisth TB, Taylor AK, Andalibi A, Svenson KL, Lusis AJ. Identification of a zinc finger protein that binds to the sterol regulatory element. Science. 1989; 245:640–643. [PubMed: 2562787]

Salehi-Ashtiani K, Luptak A, Litovchick A, Szostak JW. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. Science. 2006; 313:1788–1792. [PubMed: 16990549]

Sassanfar M, Szostak JW. An RNA motif that binds ATP. Nature. 1993; 364:550–553. [PubMed: 7687750]

Seenisamy J, Rezler EM, Powell TJ, Tye D, Gokhale V, Joshi CS, Siddiqui-Jain A, Hurley LH. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. J. Am. Chem. Soc. 2004; 126:8702–8709. [PubMed: 15250722]

Shu D, Guo P. A viral RNA that binds ATP and contains a motif similar to an ATP-binding aptamer from SELEX. J. Biol. Chem. 2003; 278:7119–7125. [PubMed: 12444088]

Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. USA. 2002; 99:11593–11598. [PubMed: 12195017]

Smirnov I, Shafer RH. Lead is unusually effective in sequence-specific folding of DNA. J. Mol. Biol. 2000; 296:1–5. [PubMed: 10656813]

Teixeira A, Tahiri-Alaoui A, West S, Thomas B, Ramadass A, Martianov I, Dye M, James W, Proudfoot NJ, Akoulitchev A. Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. Nature. 2004; 432:526–530. [PubMed: 15565159]

Todd AK, Johnston M, Neidle S. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. 2005; 33:2901–2907. [PubMed: 15914666]

Tucker BJ, Breaker RR. Riboswitches as versatile gene control elements. Curr. Opin. Struct. Biol. 2005; 15:342–348. [PubMed: 15919195]

Vu MM, Jameson NE, Masuda SJ, Lin D, Larralde-Ridaura R, Luptak A. Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. Chem. Biol. 2012; 19:1247–1254. [PubMed: 23102219]

Watson PY, Fedor MJ. The ydaO motif is an ATP-sensing riboswitch in Bacillus subtilis. Nat. Chem. Biol. 2012; 8:963–965. [PubMed: 23086297]

Wei C, Tang Q, Li C. Structural transition from the random coil to quadruplex of AG(3)(T(2)AG(3))(3) induced by Zn(2+). Biophys. Chem. 2008; 132:110–113. [PubMed: 18031921]

Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature. 2002; 419:952–956. [PubMed: 12410317]

Woodland HR, Pestell RQ. Determination of the nucleoside triphosphate contents of eggs and oocytes of Xenopus laevis. Biochem. J. 1972; 127:597–605. [PubMed: 4672799]

Zhang J, Lau MW, Ferre-D'Amare AR. Ribozymes and riboswitches: modulation of RNA function by small molecules. Biochemistry. 2010; 49:9123–9131. [PubMed: 20931966]

Zimmermann B, Bilusic I, Lorenz C, Schroeder R. Genomic SELEX: a discovery tool for genomic aptamers. Methods. 2010; 52:125–132. [PubMed: 20541015]

Highlights

- GTP aptamers are widespread in eukaryotic RNA and DNA sequences.
- Members of diverse classes of G-quadruplex regulatory elements bind GTP.
- Sequence constraints of regulatory G-quadruplexes parallel those of GTP binding.

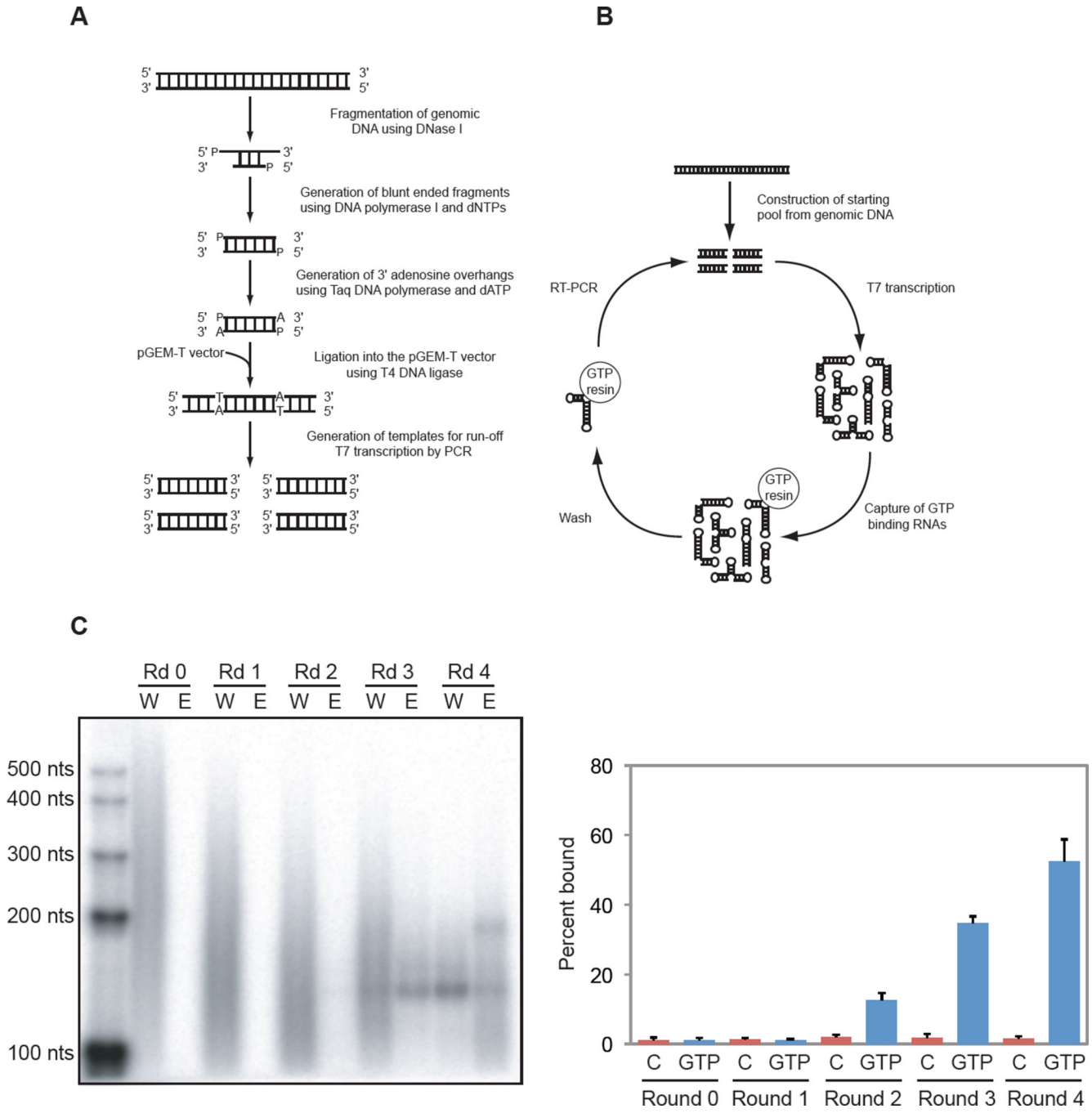

Figure 1.

Isolation of naturally occurring GTP aptamers using *in vitro* selection. a) Construction of pools of genome-encoded RNA fragments suitable for *in vitro* selection. b) *In vitro* selection of genome-encoded GTP aptamers. c) Progress of the selection for naturally occurring GTP aptamers. Left panel: 5% polyacrylamide gel showing GTP-binding activity of pool RNA from each round of the selection. W = wash fraction and E = elution fraction. Right panel: graph showing GTP-binding activity of pool RNA from each round of the selection. Red bars = pool binding to control agarose lacking GTP; blue bars = pool binding to GTP agarose. Reported values indicate the average of three independent experiments, and error

bars indicate one standard deviation. See the Supplementary Experimental Methods and Table S1 for additional experimental details.

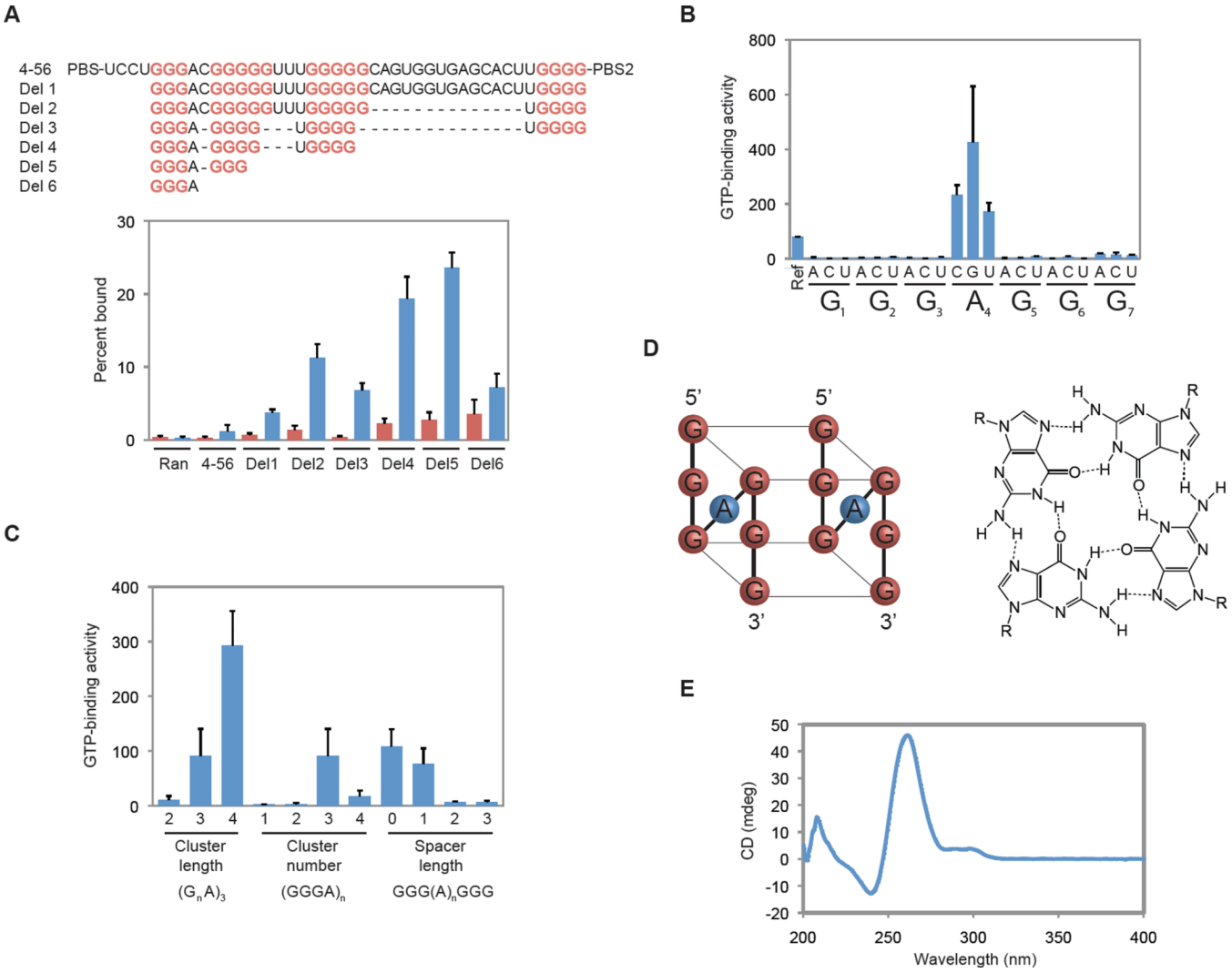

Figure 2.

Sequence requirements of the G motif aptamer. a) Minimization of the G motif. Red bars = binding to control agarose; blue bars = binding to GTP agarose. Ran = $GG(N)_{48}$. b) Characterization of all single mutants of the minimized G motif aptamer. Ref = GGGAGGG, the reference sequence. c) GTP-binding activity of the G motif aptamer as a function of G cluster size, G cluster number, and spacer length. d) Proposed G-quadruplex structure of the G motif aptamer. Left: a hypothetical G-quadruplex formed by two molecules of the minimized G motif sequence GGGAGGG. An parallel G-quadruplex is shown; note that G-quadruplexes with other topologies can also form. Right: hydrogen-bonding pattern of a G-tetrad within a G-quadruplex structure. (e) CD spectra of the G motif variant GGGGGAGGGGUGGG. For (b) and (c) GTP-binding activity = (amount of GTP bound by aptamer in a gel filtration assay) / (amount of GTP bound by a random sequence control RNA pool in the same assay). For (a), (b) and (c), reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods and Tables S1 and S2 for additional experimental details. See also Figure S1.

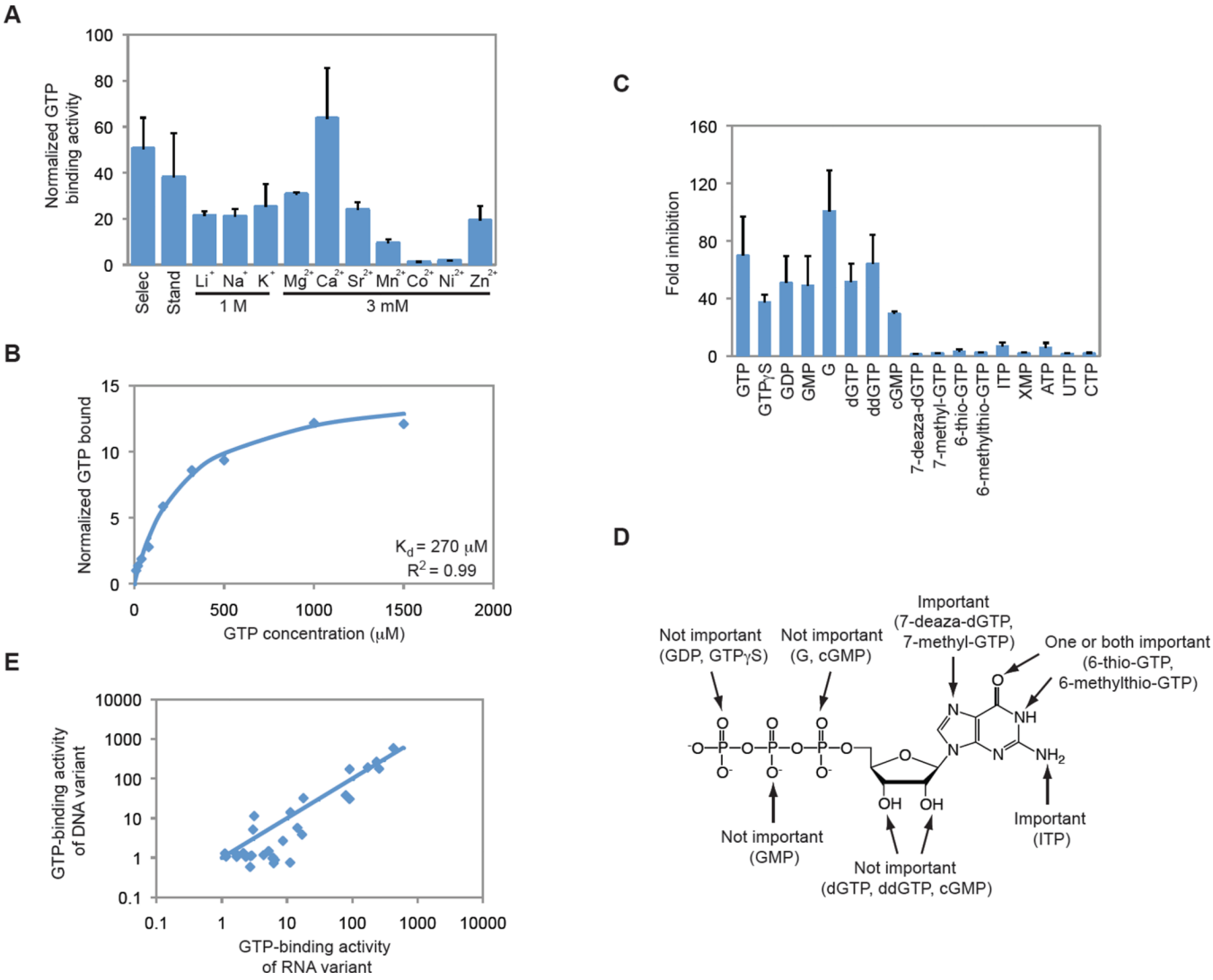

Figure 3.

Biochemical characterization of the G motif aptamer. a) Metal ion requirements of the G motif aptamer. Selec = aptamer selection buffer (20 mM $MgCl_2$, 200 mM KCl, 20 mM HEPES pH 7.1); Stand = low magnesium aptamer selection buffer (20 mM $MgCl_2$, 200 mM KCl, 20 mM HEPES pH 7.1); Monovalent buffers = 1 M monovalent metal ion, 20 mM HEPES pH 7.1; Divalent buffers = 3 mM divalent metal ion, 20 mM HEPES pH 7.1. Normalized GTP-binding activity = (amount of GTP bound in the indicated buffer as measured by gel filtration) / (amount of GTP bound in a buffer containing 20 mM HEPES pH 7.1 in the same assay). b) Dissociation constant of the G motif aptamer for GTP. Normalized GTP bound = amount of GTP bound at the indicated GTP concentration as measured by gel filtration) / (amount of GTP bound at the lowest GTP concentration used in the assay). c) Nucleotide-binding specificity of the G motif aptamer. Fold inhibition = (amount of radiolabeled GTP bound in the absence of the indicated unlabeled competitor as measured by gel filtration) / (amount of radiolabeled GTP bound in the presence of unlabeled competitor in the same assay). d) Data from competitive binding experiments mapped onto the chemical structure of GTP. e) Binding of DNA and RNA variants of the G motif to GTP. Sequences compared were those in Figures 2b and c (RNA variants) and Supplementary Figure 3 (DNA variants). The blue line shows the expected relationship if

the GTP-binding activities of DNA and RNA variants of each sequence tested were equal. For (a)-(c), experiments were performed using the sequence GGGGGAGGGGUGGG. Reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods for additional experimental details. See also Figure S2.

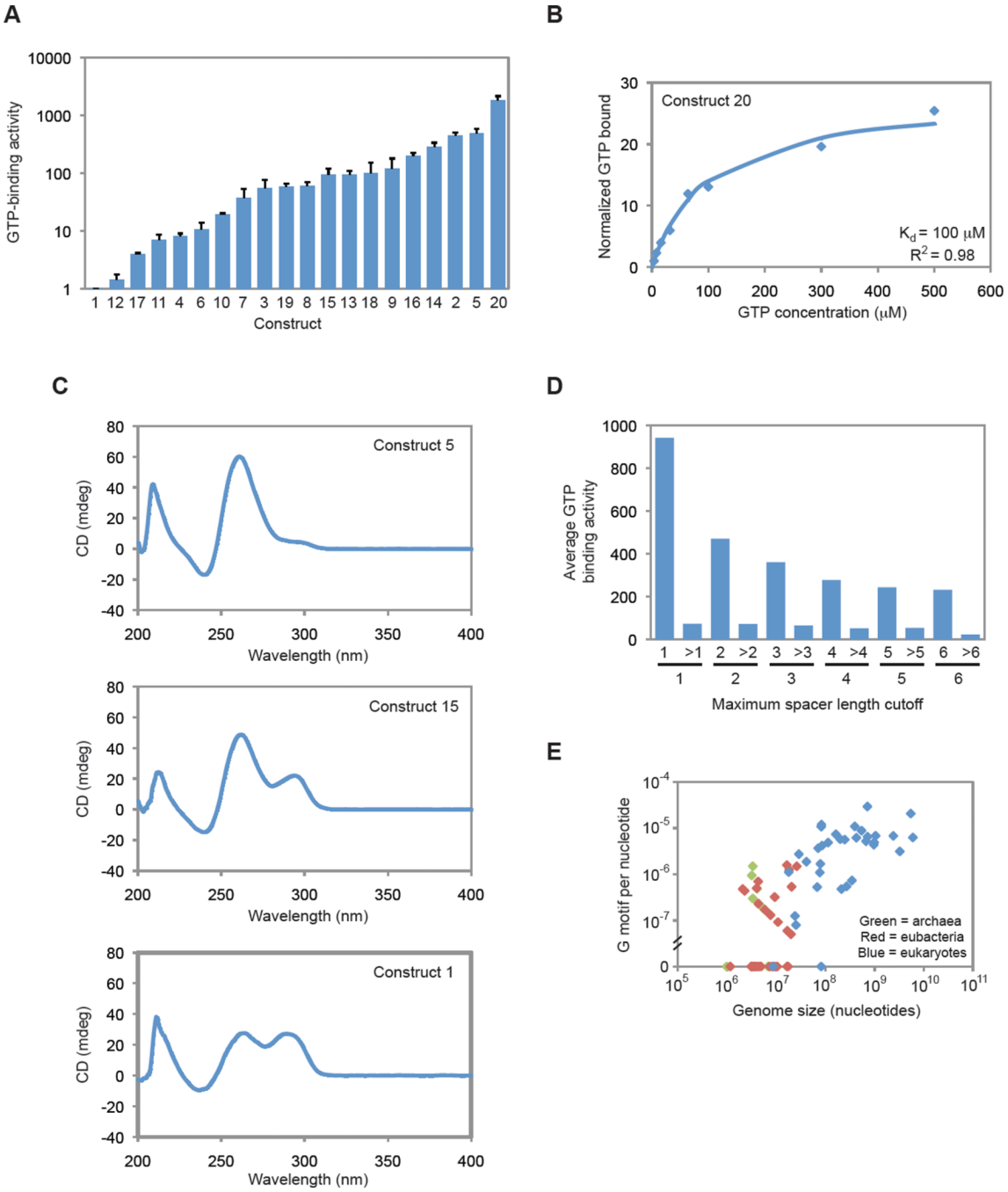

Figure 4.

Identification of the G motif in sequenced genomes. a) GTP-binding activity of 20 randomly chosen G-quadruplexes from the human genome. G-quadruplexes were identified using the quadparser algorithm. b) Dissociation constant of one of these G-quadruplexes for GTP. Normalized GTP bound = amount of GTP bound at the indicated GTP concentration as measured by gel filtration) / (amount of GTP bound at the lowest GTP concentration used in the assay). c) CD spectra of three of the human G-quadruplexes tested for the ability to bind GTP in panel (a). See Figure S3 for CD spectra of all 20 sequences. d) Average GTP-binding activity of intramolecular DNA G-quadruplexes as a function of spacer length. The height of each bar indicates the average GTP-binding activity of all G-quadruplexes

characterized in this study with the indicated maximum allowed spacer length. e) Density of the G motif in phylogenetically diverse eubacterial, archaebacterial, and eukaryotic species as a function of genome size. Examples of the G motif were identified by searching for G-quadruplexes with spacers of no more than one nucleotide using the quadparser algorithm. For (a), (b) and (d), reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods and Tables S2, S4 and S5 for additional experimental details. See also Figure S3.

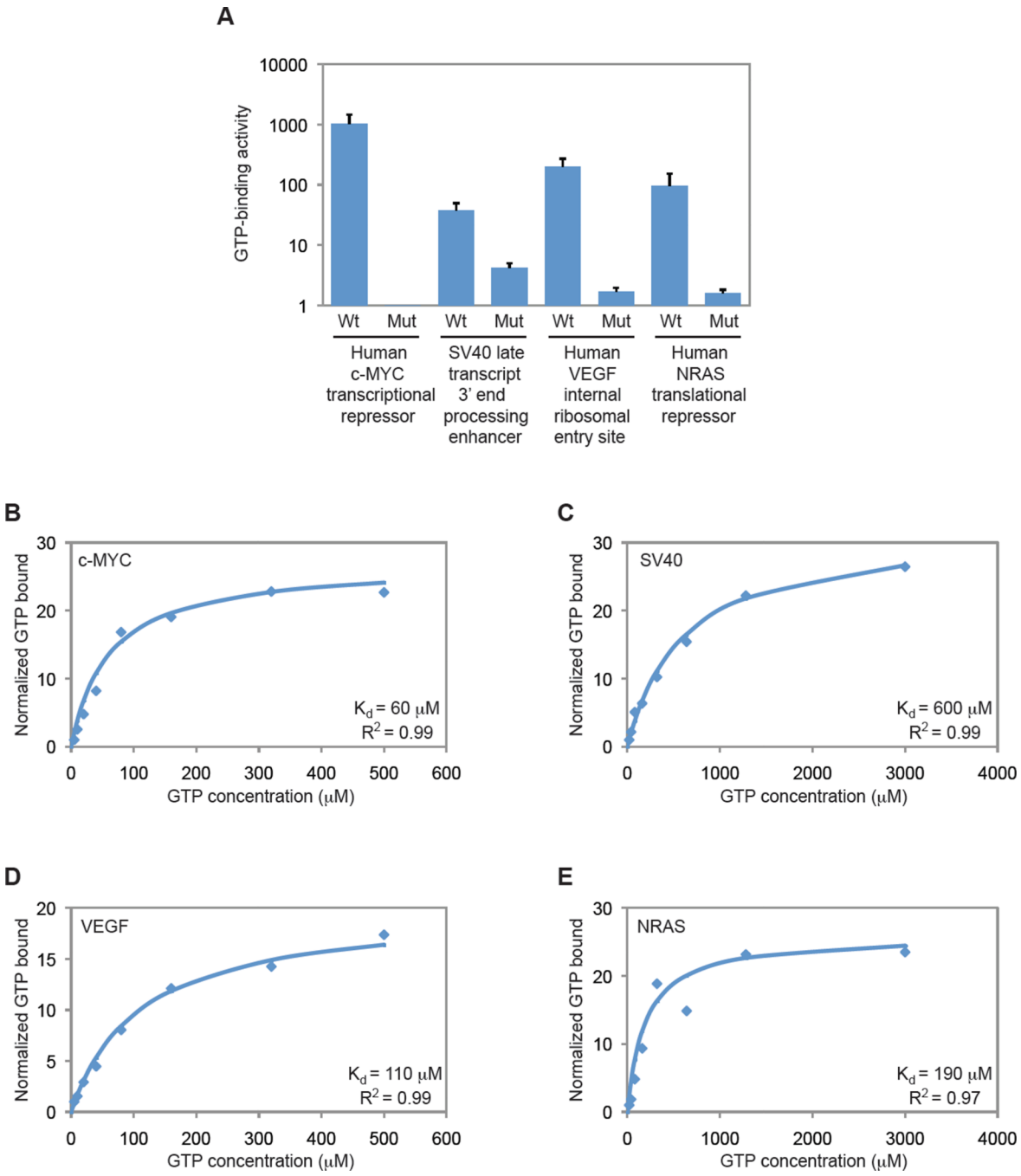

Figure 5.

GTP-binding activity of diverse classes of regulatory G-quadruplexes. a) Binding of regulatory G-quadruplexes to GTP. c-MYC transcriptional repressor wt = GGGTGGGGAGGGTGGGG, mut = GGGTAGGAAGGGTAGGA; SV40 late 3' end processing enhancer wt = GGGGGAGGUGUGGG, mut = GAGGGAGGUGUCAG; VEGF IRES wt = GGAGGAGGGGGAGGAGGA, mut = UGAUGAGUGUGAGGAGGA; NRAS translational repressor wt = GGGAGGGGCGGGUCUGGG, mut = GGUAGGGUCGGUUCUGGU. GTP-binding activity = (amount of GTP bound by aptamer in a gel filtration assay) / (amount of GTP bound by a random sequence control DNA or RNA in the same assay). b) Dissociation constant of the c-MYC transcriptional repressor for

GTP. c) Dissociation constant of the SV40 late 3' end processing enhancer for GTP. d) Dissociation constant of the VEGF internal ribosomal entry site for GTP. e) Dissociation constant of the NRAS translational repressor for GTP. For (b)-(e), normalized GTP bound = amount of GTP bound at the indicated GTP concentration as measured by gel filtration) / (amount of GTP bound at the lowest GTP concentration used in the assay). Reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods and Table S2 for additional experimental details. See also Figure S4.

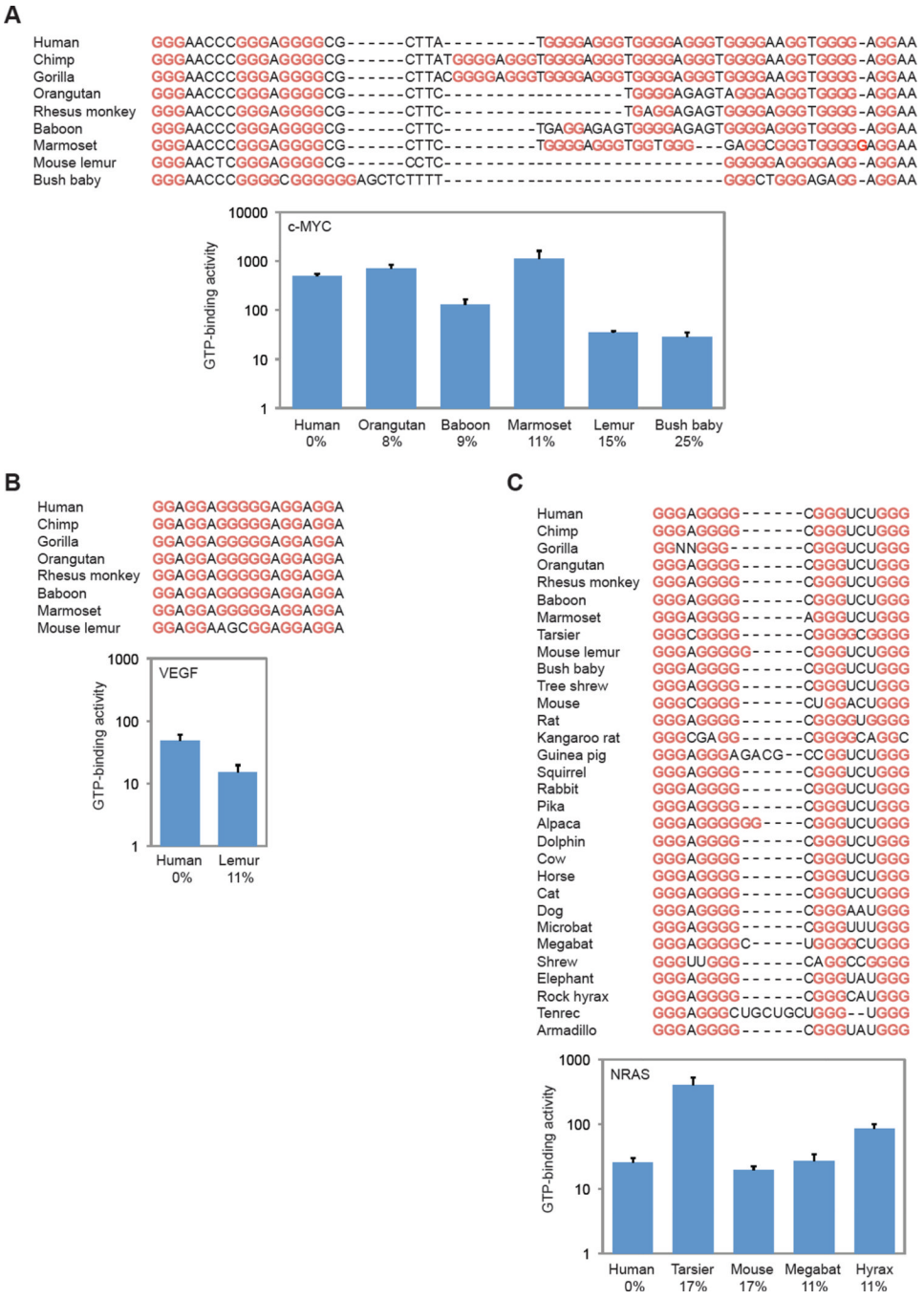

Figure 6.

Evolutionary conservation of the GTP-binding activity of G-quadruplex regulatory elements. a) Conservation of the GTP-binding activity of the c-MYC transcriptional repressor. Above: sequence alignment of variants of the c-MYC transcriptional repressor from primates, with G clusters containing two or more guanosines indicated in red. Because multiple overlapping G-quadruplexes can form in this region, between two and eleven G-quadruplexes from each species were tested for the ability to bind GTP (Table S2). Below: GTP-binding activity of six of these variants, with the percent difference between the G-rich region of each species and that of the human sequence indicated below each species name. b) Conservation of the GTP-binding activity of the VEGF IRES. Above: sequence alignment

of variants of the VEGF IRES from primates, with G clusters containing two or more guanosines indicated in red. Below: GTP-binding activity of one of these variants, with the percent difference between the sequence of this variant and that of the human variant indicated below the species name. c) Conservation of the GTP-binding activity of the NRAS translational repressor. Above: sequence alignment of variants of the NRAS translational repressor from placental mammals, with G clusters containing two or more guanosines indicated in red. Below: GTP-binding activity of five of these variants, with the percent difference between the sequence of each variant and that of the human variant indicated below each species name. Sequence alignments were obtained from the UCSC genome browser, and in some cases modified manually to highlight conservation of G clusters. Percent difference = 100 × (number of aligned positions that are different from the human sequence) / (total number of aligned positions). GTP-binding activity = (amount of GTP bound by aptamer in a gel filtration assay) / (amount of GTP bound by a random sequence control DNA or RNA in the same assay). Reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods and Table S2 for additional experimental details.

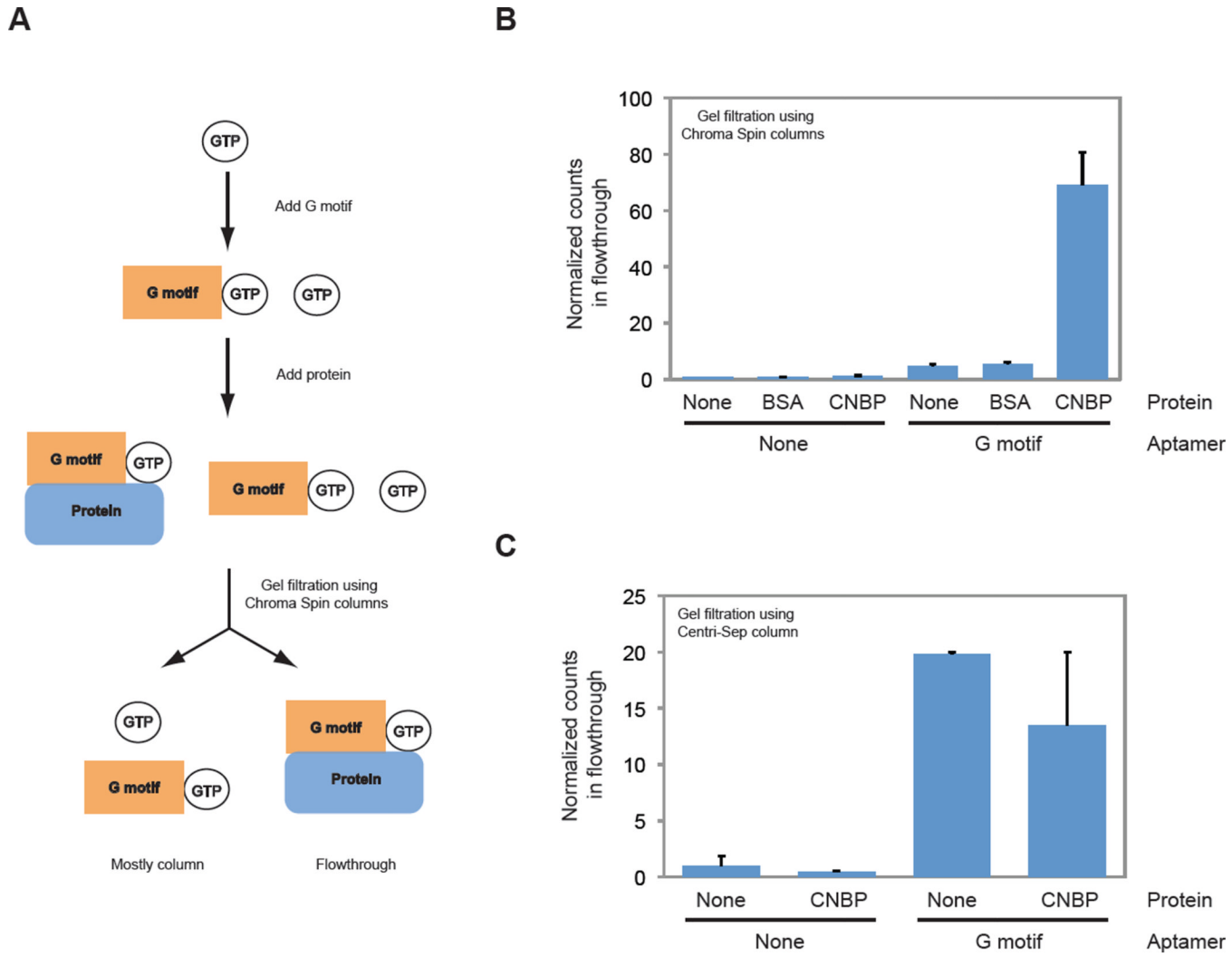

Figure 7.

CNBP protein binds G motif-GTP complexes. a) Method to assay the ability of a protein to bind G motif-GTP complexes. After mixing radiolabeled GTP with unlabeled G motif aptamer and adding the protein to be tested, G motif-GTP-protein ternary complexes are separated from G motif-GTP binary complexes and unbound GTP using Chroma Spin gel filtration columns. b) Result of this experiment using c-MYC transcriptional repressor aptamer DNA and CNBP protein. c) Result of the same experiment analyzed using Centri-Sep gel filtration columns, which separate GTP from G motif-GTP binary complexes and G motif-GTP-protein ternary complexes, but do not separate G motif-GTP binary complexes from G motif-GTP-protein ternary complexes. For (b) and (c), counts in flowthrough were normalized to the counts in the flowthrough of the minus aptamer, minus protein sample (set to a value of 1). Reported values indicate the average of three independent experiments, and error bars indicate one standard deviation. See the Supplementary Experimental Methods for additional experimental details.